МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
“ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ”
(ФГБОУ ВПО «ВГУ»)

Примеры обработки данных в пакете STATISTICA

Учебно-методическое пособие для вузов

Составители:

С.А.Ткачева

Воронеж

2014

ŋqdqɏeɎ bɆ dədɆbɎзə d ƨɔɑbɔɏ Ɉqddɏqɏè

1. Подготовительные этапы статистического исследования

Выборочный метод

Задача построения выборки возникает всякий раз, когда необходимо собрать информацию о некоторой группе или большой совокупности людей. Выборку в той или иной форме используют в ориентированных на «жесткие» статистические методы опроса. В исследованиях политических и культурных элит и даже при отборе «случаев» для включенного наблюдения и качественного анализа.

Выборка – это подмножество заданной совокупности (популяции), позволяющее делать более или менее точные выводы относительно совокупности в целом. Зачем нужно строить выборки? Прежде всего, из практических соображений, так как выборка значительно экономит средства. Проведение полномасштабной переписи или сплошного опроса населения требует значительных финансовых и трудовых затрат, которые к тому же могут оказаться напрасными, если в разработке методики исследования были допущены принципиальные просчеты. Другая причина заинтересованности в выборках связана с тем, что выборочная процедура представляет собой удобную и экономичную форму индуктивного вывода. Третья причина заключается в том, что эта процедура реализует фундаментальный принцип рандомизации, то есть случайного отбора. Поэтому наилучшей моделью отбора считается вероятностная или случайная выборка, в которой строго соблюдается принцип равенства шансов попадания в выборку для всех единиц изучаемой совокупности и для любых последовательностей таких единиц. Переписью называют процедуру сбора информации о каждом члене изучаемой группы или популяции. Все члены интересующей исследователя группы (популяции) составляют генеральную совокупность. Выборочная процедура обеспечивает обоснованность и «законность» выводов о генеральной совокупности, сделанных на основании небольшой выборки.

Выборочную ошибку определяют как расхождение между оценкой некоторого показателя, получаемого на основании исследования выборки и истинным значением этого показателя в генеральной совокупности.

Процедура построения простой случайной выборки включает в себя следующие шаги:

объектов нашего изучения, мы не можем полностью отвлечься от других шкал. Причин тому несколько.

Во-первых, соответствующие положения фактически задействованы (иногда в неявном виде) почти во всех методах анализа, в том числе и рассчитанных на номинальные данные.

Во-вторых, хотя номинальные данные являются основным предметом изучения социолога, решение большинства задач эмпирической социологии требует “увязки” процесса такого изучения с анализом данных, полученных по шкалам высоких типов. Объясняется это тем, что именно по таким шкалам измеряются столь важные для социолога характеристики респондентов, как возраст респондента, его зарплата и т.д. Поэтому строить курс анализа данных вообще без упоминания методов изучения “числовой” информации представляется нецелесообразным.

В-третьих, хотя в литературе имеется немало работ с описанием методов статистического анализа “числовых” данных, однако при этом не всегда достаточно подробно анализируются многие их аспекты, важные для социолога-практика (например, редко затрагивается проблема разбиения диапазона изменения признака на интервалы или проблема пропущенных значений). Мы постараемся ликвидировать этот пробел хотя бы для наиболее часто используемых социологом методов – вычислении мер средней тенденции и разброса для вероятностных распределений.

В социологической практике интервальность шкалы обычно сопрягается с ее *непрерывностью*, т.е. с предположением о том, что в качестве значения интервального признака в принципе может выступить любое действительное число, любая точка числовой оси.

Переходя к описанию выборочного представления функции распределения или функции плотности распределения, прежде всего отметим, что непрерывную кривую в выборочном исследовании нельзя получить никогда. Здесь мы не можем иметь, скажем, линию, похожую на известный “колокол” нормального распределения. Причина ясна: наша выборка конечна. Даже если в генеральной совокупности распределение, к примеру, нормально, а выборка - репрезентативна, мы вместо “колокола” получим лишь некоторое его подобие, составленное, например, из отрезков, соединяющих отдельные точки - полигон распределения. Заменяющая непрерывное распределение ломаная

линия может состоять также из “ступенек”, в таком случае она называется гистограммой распределения.

В математической статистике доказано, что при больших объемах выборки и достаточно мелком разбиении и гистограмма, и полигон достаточно хорошо приближают функцию плотности распределения (причем полигон делает это несколько лучше).

2. Примеры обработки данных в пакете STATISTICA

STATISTICA представляет собой интегрированную систему статистического анализа и обработки данных. Она состоит из 5 компонентов:

1) электронных таблиц для ввода и задания исходных данных, а также специальных таблиц для вывода результатов статистического анализа;
2) графической системы визуализации данных и результатов статистического анализа;
3) набора статистических модулей, в которых собраны группы логически связанных между собой статистических процедур;
4) специального инструментария для подготовки отчетов;
5) встроенных языков программирования, позволяющих расширить стандартные возможности системы.

В любом конкретном модуле можно выполнить определенный способ статистической обработки, не обращаясь к процедурам других модулей. Переключаться между модулями можно как между обычными Windows-приложениями, выбирая их на панели переключателей модулей щелчком *мыши*.

1. Инструменты для работы с данными

Данные в STATISTICA организованы в виде электронной таблицы. Таблица с исходными данными является одним из типов документов в системе STATISTICA (таблицы хранятся в файлах с расширением *.sta). Каждый тип документа выводится в своем окне в рабочей области системы. Как только окно становится активным, изменяется панель инструментов и меню. В нем появляются команды, доступные только для этого типа документов.

2. Структура электронной таблицы

Исходные данные организованы в виде таблицы. Электронная таблица состоит из строк и столбцов. Столбцы называются ***Variables*** (переменные), а строки ***Cases*** (случаи, наблюдения). Каждая переменная имеет свое имя,

формат и другие атрибуты, задаваемые пользователем. Результаты наблюдений записываются в строках таблицы. Нулевой столбец (по умолчанию содержит номера наблюдений) может содержать имена случаев. Электронная таблица с исходными данными в STATISTICA называется *Spreadsheet.* Для удобства работы с переменными, принимающими текстовые значения реализован механизм двойной записи. Каждому текстовому значению переменной ставится в соответствие некоторое числовое значение. Может быть установлено автоматически или определено пользователем. При работе с данными всегда можно переключиться с текстовой на числовую форму просмотра исходных данных. ***Data*→ Text Labels Editor** (редактор текста ярлыков) в поле **Text Label** вводить текстовое значение, в поле Numeric – численное значение (например: 0 или 1), в поле Description(описание) вводится пояснительный текст.

При работе с реальными данными часто приходится иметь дело с ситуациями, когда часть данных не была по каким-либо причинам измерена. В этом случае в соответствующую ячейку электронной таблицы не вноситься никакое значение. Ячейка остается пустой. **Однако при внутреннем хранении данных STATISTICA приписывает всем пустым ячейкам – пропущенным наблюдениям данных специальный код Missing Data Code (код пропущенных данных). Код пропущенных данных устанавливается в спецификации переменной. Значение этого кода по умолчанию равно -9999.** Пользователь всегда может установить другое значение этого кода для каждой конкретной переменной. Способ, которым пропущенные данные обрабатываются при статистическом анализе, может корректироваться индивидуально для каждого вида анализа. Обычно он может быть установлен из стартовой панели каждого модуля. Пользователь может устранить данные из вычислений, заменив их средним значением или интерполировать их. Для замены данных их средними значениями в меню **Data** выбрать команду Replase Missing Data (замена пропущенных данных на средние).

Создание файла данных

Пример 1. Создать файл Gemat.sta 6v*15c с результатами воздействия лекарства «каптоприл» на кровяное давление. Исходные данные содержатся в таблице.

Таблица. Кровяное давление (в мм. ртутного столба) до и после приема каптоприла